

Sets Have Simple Members.*

Leonid A. Levin
Boston University[†]

Abstract

The combined universal probability $\mathbf{M}(D)$ of strings x in sets D is close to $\max_{x \in D} \mathbf{M}(\{x\})$: their \sim logs differ by at most D 's information $j = \mathbf{I}(D : \mathcal{H})$ about the halting sequence \mathcal{H} . Thus if all x have complexity $\mathbf{K}(x) \geq k$, D carries $\geq i$ bits of information on each x where $i + j \sim k$. Note, there are no ways (whether natural or artificial) to generate D with significant $\mathbf{I}(D : \mathcal{H})$.

1 Introduction.

Many intellectual and computing tasks require guessing the hidden part of the environment from available observations. In different fields these tasks have various names, such as Inductive Inference, Extrapolation, Passive Learning, etc. The relevant part of the environment can be represented as an, often huge, string $x \in \{0, 1\}^*$. The known observations restrict it to a set $D \ni x$. D is typically enormous, and many situations allow replacing it with a much more concise theory representing the relevant part of what is known about x . Yet, such approaches are *ad hoc* and secondary: raw observations are anyway their ultimate source.

One popular approach to guessing, the “Occam Razor,” tells to focus on the simplest members of D . (In words, attributed to A. Einstein, “A conjecture should be made as simple as it can be, but no simpler.”) Its implementations vary: if two objects are close in simplicity, there may be legitimate disagreements of which is slightly simpler. This ambiguity is reflected in formalization of “simplicity” via the Kolmogorov Complexity function $\mathbf{K}(x)$ - the length of the shortest prefix program¹ generating x : \mathbf{K} is defined only up to an additive constant depending of the programming language. This constant is small compared to the usually huge whole bit-length of x . More mysterious is the justification of this Occam Razor principle.

A more revealing philosophy is based on the idea of “Prior”. It assumes the guessing of $x \in D$ is done by restricting to D an *a priori* probability distribution on $\{0, 1\}^*$. Again, subjective differences are reflected in ignoring moderate factors: say in asymptotic terms, priors different by a $\theta(1)$ factors are treated as equivalent. The less we know about x (before observations restricting x to D) the more “spread” is the prior, i.e. the smaller would be the variety of sets that can be ignored due to their negligible probability. This means that distributions truly prior to any knowledge, would be the largest up to $\theta(1)$ factors. Among enumerable (i.e. generatable as outputs of randomized algorithms) distributions, such largest prior does in fact exist and is $\mathbf{M}(\{x\}) = 2^{-\mathbf{K}(x)}$.

*This research was partially supported by NSF grant CCF-1049505.

[†]Computer Science department, 111 Cummings Mall, Boston, MA 02215.

¹This analysis ignores issues of finding short programs efficiently. Limited-space versions of absolute complexity results are usually straightforward. Time-limited versions often are not, due to difficulties of inverting one-way functions. However the inversion problems have time-optimal algorithms. See such discussions in [Levin 13a].

These ideas developed in [Solomonoff 64] and many subsequent papers do remove some mystery from the Occam Razor principle. Yet, they immediately yield a reservation: the simplest objects have **each** the highest universal probability, but it may still be negligible compared to the **combined** probability of complicated objects in D . This suggests that the general inference situation may be much more obscure than the widely believed Occam Razor principle describes it.

The present paper shows this could not happen, except as a purely mathematical construction. Any such D has high information $\mathbf{I}(D : \mathcal{H})$ about Halting Problem \mathcal{H} (“Turing’s Password” :-). So, they are “exotic”: there are no ways to generate such D ; see this informational version of Church-Turing Thesis discussed at the end of [Levin 13].

Consider finite sets D containing only strings of high ($\gtrsim k$) complexity. One way to find such D is to generate at random a small number of strings $x \in \{0, 1\}^k$. With a little luck, all x would have high complexity, but D would contain virtually all information about each of them.

Another (less realistic :-) method is to gain access to the halting problem sequence \mathcal{H} and use it to select for D strings x of complexity $\sim k$ from among all k -bit strings.

Then D contains little information about most of its x but much information about \mathcal{H} !

Yet another way is to combine both methods. Let v_h be the set of all strings vs with $\mathbf{K}(vs) \sim \|vs\| = \|v\| + h$. Then $\mathbf{K}(x) \sim i + h$, $\mathbf{I}(D : x) \sim i$, and $\mathbf{I}(D : \mathcal{H}) \sim h$ for most i -bit v and $x \in D = v_h$. We will see no D can be better: they all contain strings of complexity $\lesssim \min_{x \in D} \mathbf{I}(D : x) + \mathbf{I}(D : \mathcal{H})$.

The result is a follow-up to Theorem 2 in [Vereshchagin, Vitányi 10]. [Vereshchagin, Vitányi 04] provides in Appendix I more history of the concepts used here; [Kolmogorov 65, Solomonoff 64, Li, Vitányi 08] give more material on Algorithmic Information Theory. The main idea of this work is due to S. Epstein, appearing in [Epstein, Betke 11]. He is a co-author of an earlier preprint [Epstein, Levin 12] of the results below and a sole author of their many extensions in [Epstein 13].

2 Conventions and Kolmogorov Complexity Tools.

$\|x\| \stackrel{\text{def}}{=} n$ for $x \in \{0, 1\}^n$; for $a \in \mathbb{R}^+$, $\|a\| \stackrel{\text{def}}{=} \lceil |\log a| \rceil$. $\mathbf{S} \stackrel{\text{def}}{=} \{0, 1\}^*$. $(p0^-) = (p1^-) \stackrel{\text{def}}{=} p$; (\emptyset^-) is undefined. $[A] \stackrel{\text{def}}{=} 1$ if statement A holds, else $[A] \stackrel{\text{def}}{=} 0$. $\prec f$, $\succ f$, $\asymp f$, and $\lesssim f$, $\gtrsim f$, $\sim f$ denote $< f + O(1)$, $> f - O(1)$, $= f \pm O(1)$, and $< f + O(\|f+1\|)$, $> f - O(\|f+1\|)$, $= f \pm O(\|f+1\|)$, respectively.

$Q(G)$ is the probability of a set G or **mean** $\sum_x Q(\{x\})G(x)$ of a function G by a distribution Q .

We use a **prefix** algorithm U : $U(p) = x$ iff $U(p0) = U(p1) = x$. Auxiliary inputs y in U_y are not so restricted. p is **total** if U halts on all k -bit ps for some k . Our U is **universal**, i.e. minimizes (up to \asymp) complexities \mathbf{K} , $\|\mathbf{M}\|$ below, and **left-total**: if $U(p1s)$ halts, $p0$ is total.² $\mathcal{H}(i) \stackrel{\text{def}}{=} [U(i) \text{ halts}]$. All results, of course, remain valid if relativized by giving U an extra auxiliary input.

Complexity $\mathbf{K}(x|y)$ is $\min_p \{\|p\| : U_y(p) = x\}$. $\mathbf{M}_v(G) = \sum_p 2^{-\|p\|} [U(v(p^-)) \neq U(vp) \in G]$ is **universal probability**. We omit empty $|y, v$. $\|\mathbf{M}(\{x\})\| \asymp \mathbf{K}(x)$. $\mathbf{I}(x : y) \stackrel{\text{def}}{=} \mathbf{K}(x) + \mathbf{K}(y) - \mathbf{K}(x, y) \asymp \mathbf{K}(x) - \mathbf{K}(x|(y, \mathbf{K}(y)))$ is **information**. $\mathbf{I}(x : \mathcal{H}) \stackrel{\text{def}}{=} \mathbf{K}(x) - \mathbf{K}(x|\mathcal{H})$. **Non-randomness** $\mathbf{d}(x|Q, v)$ is $\lfloor |\log Q(\{x\})| \rfloor - \mathbf{K}(x|v)$. $t(x) = 2^{d(x|Q, v)}$ is a **Q-test** for any Q, v i.e. $Q(t) \leq 1$. $\lambda(d) \stackrel{\text{def}}{=} \|d\| + \mathbf{K}(\|d\|)$.

² U' is turned into left-total U by enumerating p in order of convergence of $U'(p)$ and assigning them consecutive intervals $i_p\{0, 1\}^{\mathbb{N}}$, $\|i_p\| = \|p\| + 1$ shared by p, q with $\|p\| = \|q\|$, $U'(p) = U'(q)$; then $U(p') \stackrel{\text{def}}{=} U'(p)$ if $p' \in i_p\{0, 1\}^*$.

3 The Results.

For $f(n) \in O(n)$, we use a slice $\chi_f(a) \stackrel{\text{def}}{=} \min_{v, Q=U(v)} (\|v\| + f(\mathbf{d}(a|Q, v)))$ of Kolmogorov structure function, requiring $Q(\mathbf{S})=1$ unlike [Shen 83]. $\chi \stackrel{\text{def}}{=} \chi_f$ for $f=\lambda$. Low- χ (i.e. random under simple distributions) a , Kolmogorov called **stochastic**. The other a are “exotic,” i.e. have high $\mathbf{I}(a : \mathcal{H})$:

Proposition 1 $\mathbf{I}(a : \mathcal{H}) \gtrsim \chi_f(a)$.

Proof. Let $U(vw)=a$, $\|vw\|=\mathbf{K}(a)$, v be total, v^- be not. Using $Q(a) \stackrel{\text{def}}{=} \mathbf{M}_v(a)$, and $\|v\|+\|w\| \asymp \mathbf{K}(a) \prec \mathbf{K}(a|v) + \mathbf{K}(\|v\|) + \|v\|$, we get $\|w\| - \mathbf{K}(a|v) \prec \mathbf{K}(\|v\|)$, so $\chi_f(a) \prec \lambda(v) + O(\|w\| - \mathbf{K}(a|v)) \prec \lambda(v) + O(\mathbf{K}(\|v\|))$. Now, $\mathbf{K}(a|\mathcal{H}) \prec \mathbf{K}(\|v\|) + \|w\|$, so $\mathbf{I}(a : \mathcal{H}) \succ \|v\| - \mathbf{K}(\|v\|) \gtrsim \chi_f(a)$. ■

Then we prove that all stochastic sets have simple (high \mathbf{M}) members:

Main Lemma 1 $\|\max_{x \in D} \mathbf{M}(\{x\})\| \prec \lambda(\mathbf{M}(D)) + \|\mathbf{K}(\|\mathbf{M}(D)\|)\| + \chi(D)$.

Informal outline of the proof: We break inputs of U into $\sim \mathbf{M}(D)/\mathbf{d}(D|Q, v)$ -wide intervals $p\mathbf{S}$. In each interval with total p we select one output $L_p=U(pp')$ and update a Q -test $\mathbf{t}(X)$ ($=\mathbf{t}_p(X)$). Here $(\ln \mathbf{t}(X))$ accumulates $\mathbf{M}_p(X)$, until $L=\{L_r|r < p\}$ intersects X upon which $\mathbf{t}(X)$ drops to 0. $\mathbf{t}(X)$ stops changing if its \ln exceeds $\sim \mathbf{d}(D|Q, v)$, so the restriction of $\max_p \mathbf{t}_p(X)$ to these high values (or 0) is lower-enumerable. L_p is selected to keep the mean $Q(\mathbf{t}) \leq 1$. This is possible since mean choice of L_p does not increase $Q(\mathbf{t})$, and the **minimal increase cannot exceed the mean**: this is the **key point** of the proof. At the end, small size of L limits complexity of its members, and high $\mathbf{t}(X)$ for $X \subset \mathbf{S} \setminus L$ with $\mathbf{M}(X) \geq \mathbf{M}(D)$ assures $\mathbf{d}(X|Q, v) > \mathbf{d}(D|Q, v)$, so $X \neq D$.

Formal proof: Let $v, Q=U(v)$ minimize $\chi(D)$. Given i, j , we build inductively a list $\{L_p \in U(p\mathbf{S})\}$ indexed by all total $p \in \{0, 1\}^{i+j}$. From $\{L_r|r < p\}$ we define Q -tests $\mathbf{t}_p^L(X)$: $\mathbf{t}_{0^{i+j}}^L(X) \stackrel{\text{def}}{=} 1$; $\mathbf{t}_{p+1}^L(X) \stackrel{\text{def}}{=} \mathbf{t}_p^L(X)$ if $\lceil \ln \mathbf{t}_p^L(X) \rceil \geq 2^j$ or p is not total; else $\mathbf{t}_{p+1}^L(X) \stackrel{\text{def}}{=} \mathbf{t}_p^L(X) [L_p \notin X] \exp(\mathbf{M}_p(X))$.

Let $L_{p,s}$ be $\{L_r|r < p\}$ with added $L_p=s$. Then either $\forall s \mathbf{t}_{p+1}^{L_{p,s}}(X) = \mathbf{t}_p^L(X)$ or by $(1-a) \exp(a) \leq 1$ for $a=\mathbf{M}_p(X)$ we get $\sum_s \mathbf{M}_p(\{s\}) \mathbf{t}_{p+1}^{L_{p,s}}(X) = (1-\mathbf{M}_p(X)) \exp(\mathbf{M}_p(X)) \mathbf{t}_p^L(X) \leq \mathbf{t}_p^L(X)$. So the mean $\sum_s \mathbf{M}_p(\{s\}) Q(\mathbf{t}_{p+1}^{L_{p,s}}) \leq Q(\mathbf{t}_p^L)$; thus $Q(\mathbf{t}_{p+1}^{L_{p,s}}) \leq Q(\mathbf{t}_p^L)$ for some $s \in U(p\mathbf{S})$. Such choices of $L_p=s$ assure $Q(\mathbf{t}_p^L) \leq 1$ for all $p \in \{0, 1\}^{i+j}$. Let $N \stackrel{\text{def}}{=} \exp(2^j - 1)$, Q -test $\mathbf{t}(X) \stackrel{\text{def}}{=} \max_p (\mathbf{t}_p^L(X) [N < \mathbf{t}_p^L(X)])$.

L and $\mathbf{t}(X)$ are enumerable from v, i, j . Take $i \stackrel{\text{def}}{=} \|\mathbf{M}(D)\|$, $d \stackrel{\text{def}}{=} \mathbf{d}(D|Q, v)$, $j \asymp \|d + \mathbf{K}(i|v)\|$. Then some $s \in L \cap D$, as otherwise $\mathbf{t}(D) \geq N$ and $\mathbf{d}(D|Q, v) > \mathbf{d}(D|Q, (v, i, j)) - \mathbf{K}((i, d)|v) - O(1) > \|N\| - \|d\| - \mathbf{K}(i|v) - O(1) > d$. And as $s \in L$, $\mathbf{K}(s) \prec i + j + \mathbf{K}(i, j, v) \prec i + \mathbf{K}(i) + \|\mathbf{K}(i)\| + \chi(D)$. ■

Theorem 1 $\min_{x \in D} \mathbf{K}(x) \asymp \|\max_{x \in D} \mathbf{M}(\{x\})\| \lesssim \|\mathbf{M}(D)\| + \mathbf{I}(D : \mathcal{H}) \sim \min_{x \in D} \mathbf{I}(D : x) + \mathbf{I}(D : \mathcal{H})$.

Proof. $\mathbf{I}(D : x) \asymp \mathbf{K}(x) - \mathbf{K}(x|(D, \mathbf{K}(D))) \gtrsim [x \in D] \|\mathbf{M}(D)\| = i$. The latter is achieved by a distribution $\mu_{i,D}(\{x\}) = \mathbf{M}(\{x\}) 2^i [x \in D]$. So, the Lemma and Proposition 1 complete the proof. ■

Acknowledgments. Besides Samuel Epstein, much gratitude is due to Margrit Betke, Steve Homer, Paul Vitányi, and Sasha Shen for insightful discussions.

References

- [Epstein, Betke 11] Samuel Epstein, Margrit Betke. An information theoretic representation of agent dynamics as set intersections. *2011 Conf. on Artificial General Intelligence. Lecture Notes in AI*, v. 6830 pp. 72–81. Springer. <http://arxiv.org/abs/1107.0998v1>
- [Epstein 13] Samuel Epstein. Information and Distances. PhD Dissertation, section 4. Boston University, 2013. <http://arxiv.org/abs/1304.3872v2>
- [Epstein, Levin 12] Samuel Epstein, Leonid A. Levin. Sets Have Simple Members. An earlier preprint of this paper. 2012. <http://arxiv.org/abs/1107.1458v7>
- [Kolmogorov 65] A.N. Kolmogorov. Three Approaches to the Concept of the Amount of Information. *Probl. Inf. Transm.*, 1(1):1-7, 1965.
- [Levin 13] Leonid A. Levin. Forbidden Information. *JACM*, **60/2**, 2013. <http://arxiv.org/abs/cs/0203029>
- [Levin 13a] Leonid A. Levin. Universal Heuristics: How do humans solve “unsolvable” problems? In: *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence*. Ed.: David L. Dowe. Lecture Notes in Computer Science, 7070:53-54, 2013. Also in a report for CCR/SIGACT workshop “Visions for Theoretical Computer Science”: <http://thmatters.wordpress.com/universal-heuristics/>
- [Li, Vitányi 08] Ming Li, Paul Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, 2008.
- [Shen 83] Alexander Shen. The concept of (α, β) -stochasticity in the Kolmogorov sense, and its properties. *Soviet Math. Doklady* 28/1:295-299, 1983.
- [Solomonoff 64] R.J. Solomonoff. A Formal Theory of Inductive Inference. *Inf. and Control* 7(1):1-22, 1964.
- [Vereshchagin, Vitányi 04] Nikolai Vereshchagin, Paul Vitányi. Kolmogorov’s Structure Functions and Model Selection. 2004. *Ibid*, 50/12:3265-3290, 2004. <http://arxiv.org/abs/cs/0204037>
- [Vereshchagin, Vitányi 10] Nikolai Vereshchagin, Paul Vitányi. Rate distortion and denoising of individual data using Kolmogorov complexity. *IEEE Trans. Inf. Theory*, 56/7:3438-3454, 2010. <http://arxiv.org/abs/cs/0411014>